# Information Retrieval using Poisson Query Generation Model

**S. Vasanthakumar**

PG Student, Dept. of CSE, Arunai Engineering College, Tiruvannamalai, India
*svasanthbscmca@gmail.com*

**Abstract:** ***This paper proposes a query generation model using Poisson distribution. Most existing models use multinomial distribution and score documents based on query likelihood that was computed by a query generation probabilistic model. It can be seen that the new model and the existing multinomial models are equivalent but behave differently in smoothing methods. It is found that Poisson model has several advantages over the multinomial model like naturally accommodating "per-term smoothing" and allowing more accurate background modeling. The paper presents several type of the above described model corresponding to different methods, and evaluates them.***

**Keywords:** Language models, Poisson process, query generation.

## 1. Introduction

Language models are shown to be effective for many retrieval tasks. It is like a new type of probabilistic retrieval models. There are many variants of language models which are proposed. The query generation language model is the most popular and fundamental one. It leads to the query likelihood scoring method for ranking documents. A query 'a' and a document 'd' are given in such a model and we compute the likelihood of generating query 'q' with a model. The model is estimated based on document d (i.e.) the conditional probability p(q|d). Having basis on the likelihood of generating the query we can then rank documents.

There exists a query generation language model. They are all based on either multinomial distribution or multivariate Bernoulli distribution. Especially the popular one is the multinomial distribution. They are also shown to be quite effective. The use of multinomial distribution is heavy. It is partly due to the fact that it has been successfully used in speech recognition in which a natural choice for modeling the occurrence of a particular word in a particular position in text is multinomial distribution. Multinomial distribution has the advantage of being able to model the frequency of terms in the query in comparing with multivariate Bernoulli only models the presence and absence of query terms, thus cannot capture different frequencies of query terms. From the viewpoint of retrieval, multivariate Bernoulli also has one potential advantage over in a multinomial distribution. The probabilities of all the terms must sum to 1, making it hard to accommodate per-term smoothing. But the presence probabilities of different terms are completely independent of each other, easily accommodating per-term smoothing and weighting in a multivariate Bernoulli distribution. In a multinomial model, the term absence is also indirectly captured. It is through the constraint that all the term probabilities must sum to 1.

## 2. Methodology

The Poisson distribution, a new family of query generation models is proposed in this paper. We model the frequency of each term independently with a Poisson distribution in this new family of models. To score a document, we would first estimate a multivariate Poisson model basis on the document, and then score it based on the likelihood of the query given by the estimated Poisson model. The advantage of multinomial in modeling term frequency and the advantage of the multivariate Bernoulli in accommodating per-term smoothing are combined by the Poisson model. The Poisson distribution models term frequencies, but without the constraint that all the term probabilities must sum to 1, and similar to multivariate Bernoulli, it models each term independently, thus can easily accommodate per-term smoothing. It is similar to the multivariate Bernoulli distribution.

Smoothing is critical for this new family of models as in the existing work on multinomial language models. Several smoothing methods for Poisson model in parallel to those used for multinomial distributions are derived. Moreover we compare the corresponding retrieval models with those based on multinomial distributions. It is found that while with some smoothing methods, the new model and the multinomial model lead to exactly the same formula, with some other smoothing methods they diverge. We also find that Poisson model brings in more flexibility for smoothing. The Poisson model can naturally accommodate *per-term smoothing*. This is hard to achieve with a multinomial model without heuristic twist of the semantics of a generative model. It is a key difference. It is exploited by us that this potential advantage algorithm to develop a new term-dependent smoothing algorithm for Poisson model and show that this new smoothing algorithm can improve performance over term-independent smoothing algorithms using either Poisson or multinomial model. We can see this advantage for both one-stage and two-stage smoothing.

There is another potential advantage of the Poisson model that its corresponding background model for smoothing can be improved through using a mixture model that has a closed form formula. We show that this new background model outperforms the standard background model and reduce the sensitivity of retrieval performance to the smoothing parameter.

## 3. Query Generation with Poisson Process

In the query generation framework, a basic assumption is that a query is generated with a model estimated based on a document. In most existing work, people assume that each query word is sampled independently from a multinomial distribution. Alternatively, we assume that a query is generated by sampling the frequency of words from a series of independent Poisson processes.

### *A. The Generation Process*

Let V={$w_1$,…..,$w_n$} be a vocabulary set. Let w be a piece of text composed by an author and <c($w_1$),…..c($w_n$)> be a frequency vector representing w, where c($w_i$,w) is the frequency count of term $w_i$ in text w. In retrieval, w could be either a query or a document. We consider the frequency counts of the *n* unique terms in w as *n* different types of events, sampled from *n* independent homogeneous Poisson processes, respectively. Suppose *t* is the time period during which the author composed the text. With a homogeneous Poisson process, the frequency count of each event, i.e., the number of occurrences of $w_i$, follows a Poisson distribution with associated parameter $\lambda_i t$ where $\lambda_i$ is a rate parameter characterizing the expected number of $w_i$ in a unit time. The probability density function of such a Poisson distribution is given by

$$P(c(w_i, w) = k|\lambda_i t) = \frac{e^{-\lambda_i t}(\lambda_i t)^k}{k!}$$

Without losing generality, we set *t* to the length of the text w (people write one word in a unit time), i.e. $t = |w|$. With *n* such independent Poisson processes, each explaining the generation of one term in the vocabulary, the likelihood of w to be generated from such Poisson processes can be written as

$$p(w|\Lambda) = \prod_{i=1}^{n} p(c(w_i, w)|\Lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda_i \cdot |w|}(\lambda_i . |w|)^{c(w_i,w)}}{c(w_i, w)!}$$

where $\Lambda = \{\lambda_i, \ldots., \lambda_n\}$ and $|w| = \sum_{i=1}^{n} c(w_i, w)$.

We refer to these *n* independent Poisson processes with parameter Λ as a Poisson Language Model.

Let $D = \{d_1, \ldots., d_m\}$ be an observed set of document samples generated from the Poisson process above. The maximum likelihood estimate (MLE) of $\lambda_i$ is

$$\widehat{\lambda_i} = \frac{\sum_{d \in D} c(w_i, d)}{\sum_{d \in D} \sum_{w' \in V} c(w', d)}$$

Note that this MLE is different from the MLE for the Poisson distribution without considering the document lengths. Given a document *d*, we may estimate a Poisson language model $\Lambda_d$ using *d* as a sample. The likelihood that a query *q* is generated from the document language model $\Lambda_d$ can be written as

$$p(q|d) = \prod_{w \in V} p(c(w, q)|\Lambda_d) \qquad (1)$$

This representation is clearly different from the multinomial query generation model as (1) the likelihood includes all the terms in the vocabulary *V* , instead of only those appearing in *q*, and (2) instead of the appearance of terms, the event space of this model is the frequencies of each term.

In practice, we have the flexibility to choose the vocabulary *V*. In one extreme, we can use the vocabulary of the whole collection. However, this may bring in noise and considerable computational cost.

In the other extreme, we may focus on the terms in the query and ignore other terms, but some useful information may be lost by ignoring the non-query terms. As a compromise, we may conflate all the non-query terms as one single pseudo term. In other words, we may assume that there is exactly one "non-query term" in the vocabulary for each query. In our experiments, we adopt this "pseudo non-query term" strategy. A document can be scored with the likelihood in Equation 1. However, if a query term is unseen in the document, the MLE of the Poisson distribution would assign zero probability to the term, causing the probability of the query to be zero. As in existing language modeling approaches, the main challenge of constructing a reasonable retrieval model is to find a smoothed language model for *p*(.|*d*).

## 4. Evaluation

We analytically compared the Poisson language models and multinomial language models from the perspective of query generation and retrieval. Experiment results show that the Poisson model with per-term smoothing outperforms multinomial model, Using Poisson mixture as background model also improves the retrieval performance.

**Table 1:** P(w|q) Values

| | *Query1* | *Query2* | *Query3* |
|---|---|---|---|
| Doc1 | 0.0047 | 0.0159 | 0.0001 |
| Doc2 | 0.0798 | 0.4168 | 0.0333 |
| Doc3 | 0.7788 | 0.7788 | 0.6065 |
| Doc4 | 0.1942 | 0.0555 | 0.0108 |

## 5. Conclusions

We present a query generation language models for retrieval based on Poisson distribution. We de rive several smoothing methods including single-stage smoothing and two-stage smoothing. We compare the new models with the popular multinomial retrieval models both analytically and experimentally. Our analysis shows that while our new models and multinomial models are equivalent under

some assumptions, they are generally different with some important differences.

In particular, we show that Poisson has an advantage over multinomial in naturally accommodating per-term smoothing. We exploit this property to develop a new per-term smoothing algorithm for Poisson language models, which is shown to outperform term-independent smoothing for both Poisson and multinomial models.

Furthermore, we show that a mixture background model for Poisson can be used to improve the performance and robustness over the standard Poisson background model. Our work opens up many interesting directions for further exploration in this new family of models. Further exploring the flexibilities over multinomial language models, such as length normalization and pseudo-feedback could be good future work. It is also appealing to find robust methods to learn the per-term smoothing coefficients without additional computation cost.

## References

[1] Wikipedia
[2] H. Fang, T. Tao, and C. Zhai. A formal study of information retrieval heuristics. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pages 49-56, 2004.
[3] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pages 275-281, 1998.
[4] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In Proceedings of ACM SIGIR'01, pages 334-342, Sept 2001.